Euphytica **130:** 355–364, 2003.

Simple sequence repeats as useful resources to study transcribed genes of cotton

Sukumar Saha[1,*], Mehmet Karaca[2,**], Johnie N. Jenkins[1], Allan E. Zipf[3], O. Umesh K. Reddy[4] & Ramesh V. Kantety[5,**]

[1]*USDA-ARS, Crop Science Research Laboratory, P.O. Box 5367, Mississippi State, MS 39762, U.S.A.;* [2]*Department of Field Crops, Faculty of Agriculture, Akdeniz University, 07059 Antalya, Turkey;* [3]*Department of Plant and Soil Science, Alabama A&M University, Normal, AL 35762, U.S.A.;* [4]*Biotechnology Center, Alcorn State University, MS 39096, U.S.A.;* [5]*Department of Plant Breeding, Cornell University Ithaca, NY 14853, U.S.A.; (*Author for correspondence: e-mail: ssaha@ars.usda.gov; **The authors are to be equally credited as first author)*

Received 19 March 2002; accepted 14 November 2002

Key words: Cotton, differential display, EST, SSR

Summary

Microsatellites or Simple Sequence Repeats (SSRs) are informative molecular genetic markers in many crop species. SSRs are PCR-based, highly polymorphic, abundant, widely distributed throughout the genome and inherited in a co-dominant manner in most cases. Here we describe the presence of SSRs in cDNAs of cotton. Thirty one SSR primer pairs of 220 (∼14%) tested led to PCR amplification of discrete fragments using cotton leaf cDNA as template. Sequence analysis showed 25% of 24 randomly selected cDNA clones amplified with different SSR primer pairs contained repeat motifs. We further showed that sequences from the SSR-containing cDNAs were conserved across *G. barbadense* and *G. hirsutum*, revealing the importance of the SSR markers for comparative mapping of transcribed genes. Data mining for plant SSR-ESTs from the publicly available databases identified SSRs motifs in many plant species, including cotton, in a range of 1.1 to 4.8% of the submitted ESTs for a given species.

Introduction

Microsatellites, also known as SSRs (Simple Sequence Repeats), are one of the most variable types of tandemly repeated DNA sequences found in plants and animals. These motifs can consist of a single base pair or a small number of bases (usually ranging from 1 to 6) which are repeated several times. The repeats can be either perfect tandem repeats or interrupted by several non-repeat nucleotides or compound repeats. SSRs are considered markers of choice in genome mapping because they are 1) PCR-based markers, 2) usually co-dominant, 3) often multiallelic and hypervariable, 4) randomly dispersed throughout the genome, and 5) publically available via published flanking primer sequences (Saghai-Maroof et al., 1994). The variation in SSRs is thought to be due either to slippage of DNA polymerase during replication or unequal crossing over, resulting in differences in the copy number of the core nucleotide sequences (Yu et al., 1999).

Regions flanking SSRs are highly conserved both at intra- and inter-specific levels in tetraploid cotton. PCR primers to the flanking regions have been used to identify chromosomal location of SSRs by genetic mapping (Liu et al., 2000a). SSRs have been employed in genetic analyses of cotton due to their highly polymorphic nature (Feng et al., 1997; Saha et al., 1999; Liu et al., 2000a,b; Reddy et al., 2001). Flanking SSR primer pairs specific to *Gossypium hirsutum* L. have also been used in genome mapping of other cotton species, including *G. barbadense* L., *G. nelsonii* L. and *G. australae* L. (Liu et al., 2000a,b; Qureshi et al., 2001). Recently, such PCR-based markers

were selected as one of the core sets of framework markers for the International Cotton Genome Mapping Initiative (Brubaker et al., 2000). However, the conventional idea has been that microsatellites are primarily associated with the non-coding genomic regions in eukaryotes (Gur-Arie et al., 2000; Davierwala et al., 2000). The overall objective of this paper is to demonstrate that SSR markers can be utilized as valuable resources for analyzing transcribed genes in cotton and other plant species. We also present a comprehensive review of SSRs identified in ESTs from cotton and other plant genome databases.

Materials and methods

Plant materials

Cotton lines used to study differentially expressed genes included Texas Marker-1 (TM-1) [(*Gossypium hirsutum*), provided by Dr R.J. Kohel, USDA/ARS, College Station, TX], and reciprocal crosses of inbred CM-1-90 with PS-7 and CM-1-90 with a virescent mutant (v_7v_7) line [(*Gossypium barbadense*), provided by Dr R. Percy, USDA/ARS, Maricopa, AZ]. Use of reciprocal crosses provided a scope to test the efficiency of the method in identifying cytoplasmic-specific genes from cDNAs. The TM-1 parent was developed by single seed descent and is considered the genetic standard in Upland cotton. Seeds from these cotton lines were planted in 20 × 100 cm pots with Jiffy Mix™ and grown in the greenhouse.

RNA and DNA isolation

Bulked samples of young leaves from several 14-day-old seedlings of individual lines were used for RNA and DNA extraction as described in detail by Karaca (2001). The genomic DNA was extracted with minor modification of the DNeasy Plant Mini Kit method (Qiagen Inc., Valencia, CA). TRIZOL reagent (Life Technologies Inc., Rockville, MD) was used in the presence of liquid nitrogen for RNA extraction following the modified protocol of the manufacturers. Total RNA samples were treated with DNase I (Amersham Pharmacia Biotech, Inc., Piscataway, NJ) and then assayed by PCR using genomic DNA primers to rule out any minute genomic DNA contamination (data not shown). Total RNA (mRNA) was further purified using an RNeasy kit (QIAGEN Inc., Valencia, CA).

cDNA synthesis/library construction

cDNAs used for PCR with SSR primers were synthesized from total RNA utilizing a SMART III™ Kit (Clontech Laboratories, Inc., Palo Alto, CA). In addition, fiber-specific cDNAs were prepared from RNA of 10 days post-anthesis (dpa) fiber of 'Tamcot SP37' (*G. hirsutum* L.) using the same protocol.

PCR conditions with SSR primers

Genomic DNA (gDNA) and cDNA were compared using the same SSR-specific primer pairs. The PCR reactions were performed using 80 ng gDNA or cDNA as template, 0.15 μM each of fluorescent-labeled forward and unlabeled reverse SSR primer pairs (Research Genetics, Inc., Huntsville, AL), 0.2 mM each dNTP, 1X PCR buffer (50 mM KCl, 10 mM TRIS-HCl pH 8.5, 0.1 mg/mL gelatin), 2.5 mM $MgCl_2$, and 0.5 units Ampli*Taq* Gold DNA polymerase (Perkin-Elmer Corp., Norwalk, CA)[1] in a 25 μL total reaction volume. The forward SSR primers were fluorescently labeled commercially by Perkin-Elmer, Inc. PCR was carried out in thermal cycler with the following profile: 10 min hold at 95 °C, followed by a step-down PCR consisting of 15 sec at 95 °C for denaturation, 30 sec at 65 °C for annealing and 2 min at 72 °C for synthesis. Annealing temperatures were reduced 1 °C/cycle for nine cycles and continued for 30 additional cycles at the 56 °C annealing temperature with a final extension for 30 min at 72 °C.

Resolution of PCR amplified products

Capillary electrophoretic analysis of 'multimix' fluorescently-labeled amplified DNA markers were visualized as peaks on electropherograms using the automated ABI PRISM 310 Genetic Analyzer™ equipped with GeneScan™ analysis software (PE Applied Biosystems, Foster City, CA). After identifying samples with SSR-cDNA peaks, they were electrophoresed in 2 or 3% Metaphor™ agarose gels and visualized with ethidium bromide (Figure 1). A total of 24 randomly selected fragments were eluted from the gel using a QIAquick gel extraction kit (QIAGEN Inc., Valencia, CA) and then commercially sequenced (Research Genetics, Huntsville, AL).

Data mining and EST analysis

ESTs (739,258) of different plant species, including cotton, were analyzed from the publicly available Genbank data bases for the presence of SSRs

Table 1. Comparative analysis of amplified products from cDNAs and genomic DNAs using some selected SSR pairs

SSR Primers	Product size (cDNA) bp	Product size (gDNA) bp	Repeat motifs (gDNA[1])
MP 1059	133	210	$(CA)_{16}$
MP 3599	180, 205	180, 205	$(TC)_{15}$
MP 3279	99, 111	92, 99, 103, 111	$(AG)_{15}$
MP 2544	146	180, 204	$(AG)_{11}$
MP 2590	200	200	$(AG)_{11}$
MP 786	110	110	$(AG)_{14}$
MP 1440	238, 268	238, 268	$(AG)_{15}$
MP 2895	111	220	$(GA)_{10}$
MP 3383	210	210	$(AG)_4AC(AG)_{10}$
MP 3792	174	235	$(TG)_{21}$
MP 3449	131, 137	131, 137, 146	$(CT)_6TA(CA)_{12}$
MP 3895	175, 195	175, 195	$(TG)_{10}$
CML 60[2]	96, 121	96, 121	$(AG)_8$
CML 66[2]	125, 135	125, 135	$(AG)_{12}$
MP 252	183	183	$(CT)_{21}$
MP 3816	195, 200	200	–
MP 1414[2]	138, 156	138, 156	$(AG)_{16}$

[1] Liu et al., 2000.
[2] cDNA and genomic DNA source TM-1.
All other sample sources from F_1 (PS-7xCM-1-90).

following the method of Kantety et al. (2001). Although we identified over 18,000 SSR-ESTs in different plant species from data mining of GenBank, some of them may have been represented more frequently than others because of the random nature of EST sequencing. This redundancy was reduced by EST clustering within species using the strategy implemented in Kantety et al. (2001). The resulting consensus sequences and singletons were pooled together to create a non-redundant (NR) data set of ESTs and then searched for the presence of the SSR primer pairs by BLAST analysis. We used a minimum of 9 di-nucleotide repeats of SSR-containing ESTs in our data mining method as opposed to the SSR-specific primers which amplified 7 and higher numbers of di-nucleotide repeats.

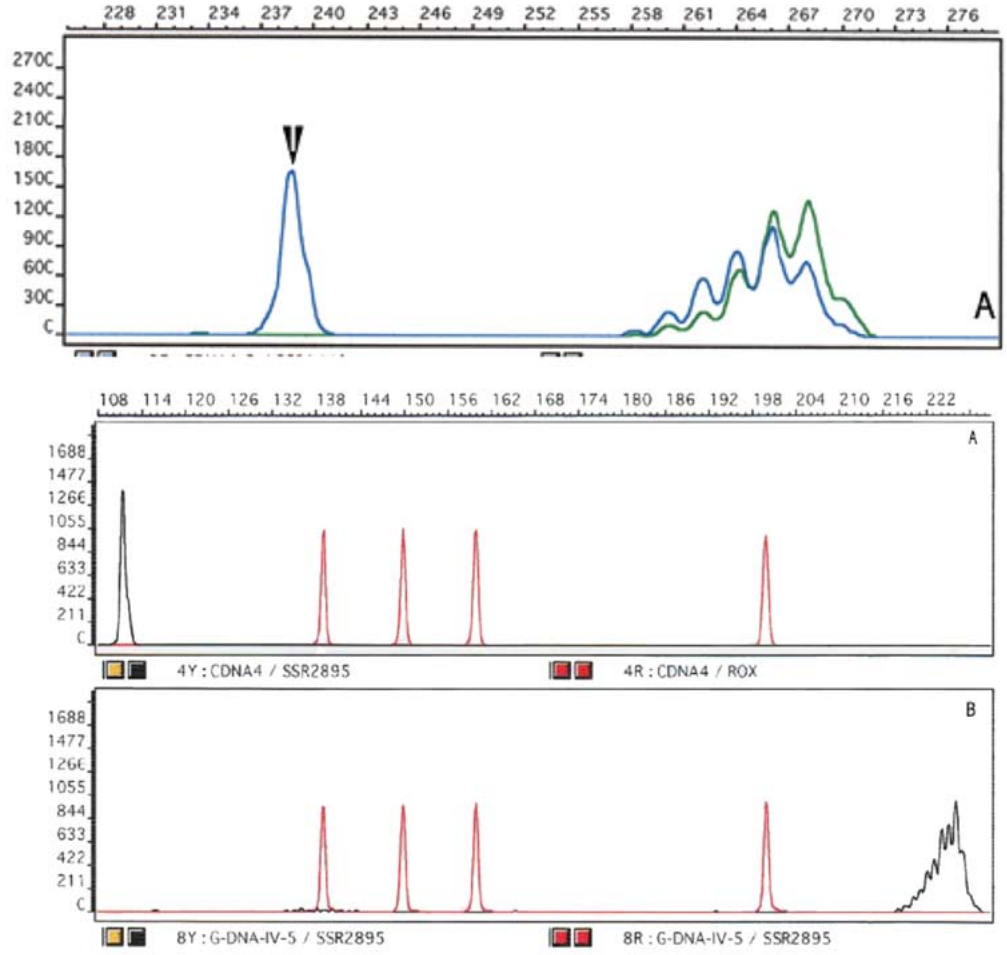
Results

Comparison of SSRs amplified from cDNAs and genomic DNA samples

SSR-specific primer pairs (ex. MP 1414) could be successfully used to distinguish differentially expressed cDNA samples (Figure 1).

A total of 220 SSR primer pairs (MapPairs™ from Research Genetics, Huntsville, AL) were initially screened using genomic DNA and cDNA from TM-1, a standard Upland cotton. Our results demonstrated that all of the SSR primer pairs under study amplified cotton genomic DNAs while only 31 (14%) could amplify the cDNAs. Many of the amplified products were of different sizes between cDNAs and genomic DNAs (Table 1, Figure 2). The number of amplified products in cDNAs ranged from 1 to 2/primer pair (average of 0.9 SSR markers/primer pair from cDNAs, and 1.5 SSR markers/primer pair from genomic DNA).

We randomly selected a total of 24 cDNA-SSR amplified fragments for sequence analysis. Analysis indicated that six of the 24 cDNA-SSR amplified fragments contained SSR motifs similar to that of genomic DNA repeats (Table 2). The banding profiles of many of the amplified products showed the presence of stutter bands typical to SSRs in many cDNAs and genomic DNAs (Figure 1A). Sequences from the amplified fragments generated using cDNA confirmed the presence of tandem repeats similar to fragments amplified using genomic DNA (Table 2).




```
                 . . . 250 . . . 260 . . . 270 . . . 280 . . . 290 . . . 300
hirsutum    241:AAGGGGAATCTTTAGGAACATGGTTGGCGACTCCAAAGAAGTGGATGTGAGCGAGCTCGA
barbadense    1:.........................GGCGACTCCAAAGAAGTGGATGTGAGCGAGCTCGA

                 . . . 310 . . . 320 . . . 330 . . . 340 . . . 350 . . . 360
hirsutum    301:GTTAACCGATCAAAGGCTTTTACTCGAACGGCTCGTTAAATCTGCCGAGGATGATCCTGA
barbadense   36:GTTAACCGATCAAAGGCTTTTACTCGAACGGCTCGTTAAATCTGCCGAGGATGATCCTGA

                 . . . 370 . . . 380 . . . 390 . . . 400 . . . 410 . . . 420
hirsutum    361:ACAGTTCTTTGATCGTATGAGGAAACGAATCGAGGCGTAATGCTGGATTTACTTACTTTT
barbadense   96:ACAGTTCTTTGATCGTATGAGGAAACGAATCGAGGCGTAATGCTGGATTTACTTACTTTT

                 . . . 430 . . . 440 . . . 450 . . . 460 . . . 470 . . . 480
hirsutum    421:TCTTTTTCTTTTTTTTTT.....TGTGTGTGTGTGTGTGTGTGTGTGTGGTTCA.AGTG
barbadense  156:TCTTTTTCTTTTTTTTTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGGCACAGTG

                 . . . 490 . . . 500 . . . 510 . . . 520 . . . 530 . . . 540
hirsutum    475:TTCTGTTAAAGTTTAAAAGAAAGAACAAGAGAACAACTCTTTGCTGGTTTTTATATTGA
barbadense  216:TTCTGTTAA.GTTTAAAAGAAAGAACAGAGAGAACCTCTTTTT.CTG............
```

.

⇐= *Figures 1-3.*

Figure 1. Cytoplasmic-specific cDNA-SSR differentially expressed transcripts using the same SSR-specific primer pair (MP 1414) between two samples of reciprocal crosses. Blue and green color represented cDNAs samples from the reciprocal crosses of F_1 plants of (PS-7 × CM-1-90) and (CM-1-90 × PS-7), respectively. Electropherograms: (A) Showing polymorphic fluorescently labeled amplified cDNA bands. The x-axis represents size of the DNA fragments and the y-axis shows the amount of the amplified products in arbitrary units; (B) Representing the same PCR products separated by agarose gel electrophoresis. Note the arrow indicating a polymorphic cytoplasmic-specific band present in one of the reciprocal F_1 crosses and absent in the other F_1.

Figure 2. Electrophoregram showing difference in banding profiles between cDNA (A) and genomic DNA (B) of the same sample using same SSR primer pair.

Figure 3. Sequence alignment of a fiber cDNA sequence from *Gossypium hirsutum* L. and *G. barbadense* L. Note presence of polymorphic dinucleotide repeat region from nucleotide position 439 to 472. Highlighted areas show base identity between the two species.

Conservation of cDNA-SSRs across cotton species

We observed that the same SSR primer pairs could be used to amplify cDNAs from leaf and fiber cells in both allotetraploid *G. hirsutum* L. (AD genome) and *G. barbadense* L. (AD genome). We identified and sequenced 200 fiber-specific cDNA clones from *G. hirsutum* L. Tamcot Sphinx (unpublished data). Within the 200 fiber-specific clones, three different dinucleotide and trinucleotide repeat motif groups were found – TG, AT and CTT, respectively (Table 2). SSR primer pairs, amplified cDNAs of *G. hirsutum*, were designed to amplify genomic DNAs of *G. barbadense* L. homologues and it was found that in some cases the repeat sizes were polymorphic between these two species. DNA sequences of the amplified product from both species revealed the presence of the same repeat motif (TG) in both homologues (Figure 3). The alignment of the cDNA sequences indicated a very high level of sequence conservation in that fiber-associated gene from both species (Figure 3).

Data mining for SSRs in plant EST-databases

We analyzed 739,258 ESTs from the publicly available databases of the major dicot and monocot crop species, including cotton, for the presence of repeat motifs that were at least 18 (di- and tri-) and 20 (tetra-) bases in length (Table 3). According to our results, many crop species contained SSRs within ESTs and their frequency ranged from 1.1% to 4.8% of the sequences in the Genbank EST database.

Discussion

Amplified fragments of the same size using both genomic and cDNAs generally indicated exon-specific segments, while larger fragments, amplified from genomic DNA, suggested the presence of an intron(s).

In addition, the MP 1414 SSR primer pair was able to identify cytoplasm-specific cDNA transcripts, based on the differences in amplification profiles of the reciprocal crosses of the *G. barbadense* lines (Figure 1). All of the SSR primer pairs in our study were originally developed from sequence analysis of the genomic DNAs. The identification of cytoplasm-specific cDNA transcripts using nuclear genome-specific SSR primers showed that there is some similarity, at least in the primer flanking regions, between the nuclear and cytoplasmic genomes in cotton. Few reports are available in cotton on genes specific to cytoplasmic genomes (Kohel & Benedict, 1971; Katterman & Endrizzi, 1973; Galau & Wilkins, 1989). Our results confirmed the importance of SSR-specific primers in transcript analysis of both the nuclear and cytoplasmic genomes.

DNA sequence results in our experiments revealed the absence of the repeat motifs in some of the cDNAs (unpublished data). However, all of the amplified products of the cDNAs retained the presence of conserved flanking primer sequences, indicating the amplification product was not an artifact. Comparing the genomic SSR fragment sequences from CottonDB with the cDNAs lacking repeat motifs showed high similarity within the non-repeat regions of the genomic fragments. Our results suggested that the repeat motifs of these cDNAs might be located in the intron(s).

We found that ∼14% (31/220) of the SSR primer pairs successfully amplified cDNA, of which at least 25% possessed repeat motifs. Cardle et al. (2000) reported that 3% of the ESTs available for cotton in EMBL (6/22/99) contained at least one repeat motif. Our percentage of SSR-containing ESTs was greater because our method was biased towards finding cDNAs containing SSRs. They observed that the average distance between SSRs for different species was 6.8 kb, which was identical to that found in *Arabidopsis*. However, the highest frequency of EST-derived

Table 2. Partial sequences of representative cDNAs showing the presence of repeat motifs in ESTs

Cotton SSR primer pairs (R/F)	cDNA Sequences	Sources	Homology
F5′-CCGAAATATACTT GTCATCTAAACG R5′-CCCCCGGAC TAATTTTTCAA	TGTCATCTAAACGAATTAGAGATTTTTCCATC ACTAATTTCAATAAATGCCAAATAAATTTTAAT ACAAAAAATAATTTAAAAAACCTTAATAAAATT TTAATAATAAA**AGAGAGAGAGAGAGAGAGAGA GAGAGAGAGAGAGAGAGAGAG**TTGAGAGA GGAATAAAAGCTGAGTGAGAAATTAAACAGG GAATTGAGAGAGATTTGAGAATGTGAGAAAG AGATGAGTTGAGATTGAC	Leaf tissues *G. barbadense*	None
F5′-AAAAACCCC TTTCCATCCAT R5′-GGGTGTCCT TCCCAAAAATT	TTTTTGGGGAAAGGTAGGTAGTTTCCCGCAA CCAAACATATACTTCACACTGTAAATCATATT GCAT**TCTCTCTCTCTCTCTCTCTCTCTCTC TC**TTATATAGAGAGAGGGGCGTATA	Leaf tissues *G. barbadense*	None
F5′-GTGGATGTGAGC GAGCTCGAG R5′-GTTAGTTTGAC GTACCTGCC	TCTGCCGAGGATGATCCTGAACAGTTCTTTG ATCGTATGAGGAAACGAATCGAGGCGTAATG CTGGATTTACTTACTTTTTCTTTTTCTTTTTTT TTTT**TGTGTGTGTGTGTGTGTGTGTGTG**G TTCAAGTGTTCTGTTAAAGTTTAAAAGAAAGA ACAAAGAGAACAA	Fiber tissues *G. hirsutum*	BNLGHi10517
F5′-CTAAGGGAAGTGGC-AAGC R5′-CCAAACAGAAC AACACTCATC	GCTCAGCTCTGAAACTGAAAGTCGACGAGGT TTTTGATGGGAATTTATAGATAATTAATTGAA TCGTTGTAGAATTCTTTCTATATTTGTGTCGT TTTGAATTGTGTAACAGCTTACAATCATAAGT GTTATTGTTCTAAGGGAAGTGGCAAGCA**ATA TATATATATAT**GCACTTTTATATGTAGCTCCA AATGATGAGTGTTGTTCTGTTTGGTAATTGTT GTAAAGTTTTAGTTTGTATCATCACTAATAAT GTTAAT	Fiber tissues *G. hirsutum*	GA_Eb0022L08f GA_Eb0013K04f GA_Eb0011F02f GA_Eb0030C14 GA_Eb0013L17f GA_Eb0015L05f GA_Eb0015J02f GA_Eb0024H08f BNLGHi6862
F5′-GAAGCAAATGC AGAGACAC R5′CTTGCTGACGAC GAGTA	ACACAAGAAATCAATCCATCAAAGAT:CAAGG AAGAACCCACAACAACCATACATATTATGGT ATG:CAATCTTGATTTGACAACATAAACATCA CCCAAGAAACATATACATATCGGTATTCACTC AAGCAGCTTCCT**CTTCTTCTT**CCTCTCCTTCT CATACTCGTCGTCAGCAAGTTGCATCCTG	Fiber tissues *G. hirsutum*	GA_Eb0009A08f GA_Ea0027K18f BNLGHi5124
F5′-GAAGATTCCA TCTGCAGACCCAG R5′-CAACAAAACCA TAAACATGAACTC	ATTTAGGAAAGACACTCATTGGGTTGAAGGG TGTCACGGCAGGGA**AGAGAGAGAGAGAGA G**AAGGTGCTTCCTGGGCA	Leaf tissues *G. hirsutum*	GA_Eb0027E04f BNLGHi6776 GA_Eb0028E07f GA_Eb0037H14f

SSRs was found in rice at 3.4 kb between SSRs, followed by soybean (7.4 kb), maize (8.1 kb), tomato (11.1 kb), poplar (14.0 kb) and cotton (20.0 kb).

SSR-containing ESTs can also be used as markers for comparative mapping and evolutionary studies between different *Gossypium* species. The alignment of the fiber-specific cDNA sequences from *G. hirsutum* and *G. barbadense* indicated a very high level of sequence conservation (Figure 3). However, some changes have occurred between these two species during evolution, including expansion or contraction of the repeat motif and insertion/deletions of single nucleotides mostly around the regions flanking the repeat motifs. We previously reported similar changes

Table 3. Analysis of ESTs from multiple species for the presence of SSR motifs that are at least 18 (di- and tri-) to 20 (tetra-) bases in length

Source	Number of ESTs	Number of SSR-ESTs	Percentage of SSRs in ESTs
Dicotyledonous species:			
Gossypium arboreum	26,630	348	1.3
Gossypium hirsutum	9,447	133	1.4
Arabidopsis thaliana	112,500	1,673	1.5
Lycopersicon esculentum	94,556	1,313	1.4
Lycopersicon hirsutum	2,504	42	1.7
Lycopersicon pennellii	2,919	40	1.4
Solanum tuberosum	20,679	228	1.1
Glycine max	137,698	3,099	2.3
Total	406,933	6,876	1.7
Monocotyledonous species:			
Triticum aestivum	54,701	2,044	3.7
Triticum monococcum	2,857	45	1.6
Triticum turgidum	2,720	130	4.8
Oryza sativa	69,164	3,291	4.8
Zea mays	76,055	1,222	1.6
Hordeum vulgare	69,086	2,418	3.5
Sorghum bicolor	48,086	1,710	3.6
Sorghum propinquum	8,455	231	2.7
Sorghum halepense	1,201	43	3.6
Total	332,325	11,134	3.4
Grand Total	739,258	18,010	2.4

in SSRs in genomic sequences from *G. hirsutum* L. (AD), *G. nelsonii* L. (G) and *G. australae* L. (G) (Qureshi et al., 2001).

Using our non-redundant Genbank EST database (NR-EST), we identified 133 SSR-containing ESTs (1.3%) from allotetraploid *G. hirsutum* L. [AD genome] and 348 SSR-containing ESTs (1.4%) from diploid *G. arboreum* L. [A genome] (Table 3). Primer pairs can be designed from these NR-EST sequences to be used directly for cDNA amplification to analyze differential gene expression.

The trinucleotide repeats were present in highest percentage among all of the SSR-containing ESTs in both *G. hirsutum* L. and *G. arboreum* L. (Table 4). In humans, the expansion of trinucleotide repeats in the coding regions was sometimes related with neurological diseases like Huntington's Disease (The Huntington's Disease collaborative Research Group, 1993). Ayers et al. (1997) reported that the *waxy* gene in rice was associated with a poly (CT) SSR in the 5′ UTR region and that the amylose content was correlated with the repeat size.

Our results showed that GA/CT among the dinucleotide repeats and AAG/TTC among the trinucleotide repeats were the most common in SSR-containing ESTs in both *G. hirsutum* L. and *G. arboreum* L. (Figure 4). Kantety et al. (2001) also observed dinucleotide motifs, GA/CT was the most abundant in ESTs among graminaceous species. They mentioned that GA/CT motif can represent GAG, AGA, UCU and CUC codons which correspond to amino acids Arg, Glu, Ala, and Leu, respectively. Ala and Leu are present in proteins at a high frequency of 8% and 10%, respectively. They concluded that this might be the reason that GA/CT motifs are present at such high frequency in many crop species. In addition, they found that the most common tri-nucleotide repeat in the majority of the graminaceous species was CGG/GCC, in contrast to our result where AAG/TTC was the most common in both the diploid and tetraploid species of cotton (Figure 3). Wren et al. (2000)

Table 4. Summary of the EST analysis from two species within the genus *Gossypium*

Source	Number of ESTs	Number of SSR-ESTs	% of ESTs with SSRs	Number of repeats			% of repeats over all SSR-ESTs		
				di-	tri-	tetra-	di-	tri-	tetra-
G. hirsutum L.	9,447	133	1.4	42	82	9	32	62	6
G. arboreum L.	26,630	348	1.3	135	199	14	39	57	4
Total	36,077	481	1.3	177	281	23	37	58	5

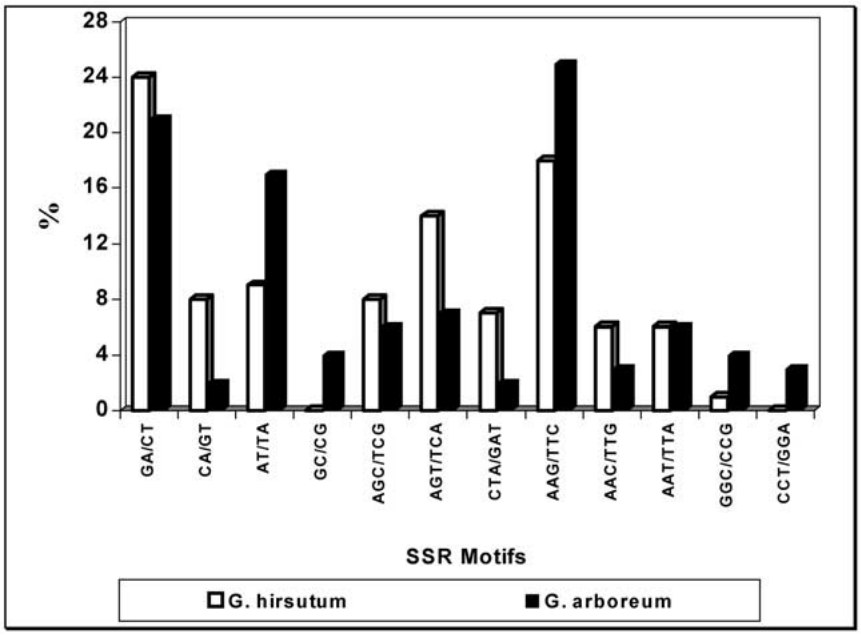


Figure 4. Distribution of SSR-ESTs in two cotton species based on the repeat motif sequences. Total number of SSR-containing ESTs in *G. hirsutum* L. and *G. arboreum* L. were 104 and 304, respectively for this analysis.

observed in humans that the 5′ and 3′ UTR were known to harbor more genetic variation than seen in coding sequences. This variation is due to two factors: the presence of repetitive sequence with regulatory functions (e.g. mRNA stability) and selection against repeat polymorphism within coding sequences. Of the 136 trinucleotide repeats identified in 5′ UTR regions, 101 of them were CGG or CCG. The 3′ UTR regions were biased towards mononucleotide repeats (Wren et al., 2000).

Allotetraploid cotton species (AD genome) belong to a 1–2 million year old lineage that reunited diploid A and D genomes and diverged 5–10 million years ago (Brubaker et al., 1999). Considering the economic importance of the genus *Gossypium,* it will be useful to further investigate the role of SSRs in the evolution of functional genes in *Gossypium*. Polymorphic SSRs in genes will provide valuable evolutionary information about selection forces acting on genes and the physiological role of functional genes in evolution.

We have a higher percentage of SSR-containing ESTs in comparison to the publicly available GeneBank data base because our experiments were specifically targeted to identify cDNAs containing SSRs using the SSR-specific primer pairs. Also, we used SSR-specific primers that could amplify 7 and higher numbers of di-nucleotide repeats of SSRs in cDNAs versus the presence of a minimum of 9 di-nucleotide repeats of SSR-containing ESTs in our data mining method.

The low number of amplified products from cDNAs provided an additional advantage compared to other differential display methods, e.g. AFLP cDNA fingerprinting (Saha et al., 2001), as the fragments

can be readily isolated both for sequencing and for analysis. However, in the same token, this method does not provide the scope to compare large numbers of differentially expressed transcripts seen by other differential display techniques.

To the best of our knowledge, there are no reports on the use of SSR-specific primers in the identification of transcribed plant genes in cotton or any other crop species. SSR-primed amplification of cDNA was useful for the identification of both nuclear and cytoplasmic genes depending on the nature of the primer pairs and samples. These results further demonstrated the merit of this method in comparative genome scanning of transcribed genes and for predictions about the role of SSRs in evolution.

Acknowledgements

Mention of trademark or proprietary product does not constitute a guarantee or warranty of the product by the United States Department of Agriculture and does not imply its approval to the exclusion of other products that may also be suitable. Part of a Ph.D. Dissertation by Mehmet Karaca at Mississippi State University. The authors are grateful to Dr Frank Callahan, USDA/ARS, MS, Dr Roy Cantrell, Cotton Incorporated and Dr Maricio Ullola, USDA/ARS, CA for assistance in reviewing the manuscript.

References

Ayers, N.M., A.M. McClung, P.D. Larkin, H.F.J. Bligh, C.A. Jones & W.D. Park, 1997. Microsatellites and a single nucleotide polymorphism differentiate apparent amylose classes in an extended pedigree of US rice germplasm. Theor Appl Genet 94: 773–781.

Brubaker, C., A.H. Paterson & J.F. Wendel, 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome 42: 184–203.

Brubaker, C., R. Cantrell, M. Giband, B. Lyon & T. Wilkins, 2000. Letter to Journal of Cotton Science community. J Cotton Sci 4: 149–151.

Cardle, L., L. Ramsy, D. Milbourne, M. Macaulay, D. Marshall & R. Waugh, 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156: 847–854.

Davierwala, A.P., W. Ramakrishna, P.K. Rajekar & V.S. Gupta, 2000. Sequence variation at a complex microsatellite locus in rice and its conservation in cereals. Theor Appl Genet 101: 1291–1298.

Diatchenko, L., Y.C. Lau, A.P. Campbell, A. Chenchik, F. Mooradam, B. Huang, S. Lukyanov, K. Lukyanov, N. Gurskaya, E.D. Sverdlov & P.D. Siebert, 1996. Suppression subtractive hybridization: A method for generating differentially regulated or time-specific cDNA probes and libraries. Proc Natl Acad Sci 93: 6025–6030.

Feng X., S. Saha & K.M. Soliman, 1997. DNA fingerprinting in cotton using AFLP. Focus 19: 11–12.

Galau, G.A. & T.A. Wilkins, 1989. Alloplasmic male sterility in AD allotetraploid *Gossypium hirsutum* upon replacement of its resident A cytoplasm with that of D species *G. harknessii*. Theor Appl Genet 78: 23–30.

Gur-Arie, R., C.J. Cohen, Y. Eitan, L. Shelef, E.M. Hallerman & Y. Kashi, 2000. Simple sequence repeats in *Escherichia coli*: Abundance, distribution, composition and polymorphism. Genome Res 10: 62–71.

Huntington's Disease Collaborative Research Group, 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's Disease chromosomes. Cell 72: 971–983.

Jones, W.S., D. Cai, O.S. Weislow & B. Esmaeli-Azad, 1997. Generation of multiple mRNA fingerprints using fluorescence-based differential display and an automated DNA sequencer. Biotechniques 22: 536–543.

Kantety, R.V., M.L. Rota, D.E. Mathews & M.E. Sorrells, 2001. Data mining for simple sequence tags from barley, maize, rice and wheat. Plant Mol Biol (in press).

Karaca, M., 2001. Characterization of *Cynodon* spp. and *Gossypium* spp. Genomes using Molecular and Cytological Techniques. Ph.D. diss. Mississippi State Univ., Mississippi State, MS.

Katterman, F.H. & J.E. Endrizzi, 1973. Studies on the 70S ribosomal content of a plastid mutant in *Gossypium hirsutum*. Plant Physiol 51: 1138–1139.

Kohel, R.J., 1978. Linkage tests in Upland cotton. III. Crop Sci 18: 844–847.

Kohel, R.J. & C.R. Benedict, 1971. Description and CO_2 metabolism of aberrant and normal chloroplasts in variegated cotton *Gossypium hirsutum* L. Crop Sci 11: 486–488.

Liang, P. & A.B. Pardee, 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257: 967–971.

Liu, S., S. Saha, D. Stelly, B. Burr & R.G. Cantrell, 2000a. Chromosomal assignment of microsatellite loci in cotton. J Heredity 91: 326–332.

Liu, S., R.G. Cantrell, J.C. McCarty, Jr. & J. McD Stewart, 2000b. Simple sequence repeat-based assessment of genetic diversity in cotton race stock accessions. Crop Sci 40: 1459–1469.

Paterson, A.H. & R.H. Smith, 1999. Future horizons: biotechnology for cotton improvement. In: C.W. Smith & J.T. Cothren (Eds.), Cotton, pp. 415–432. John Wiley & Sons, Inc., New York, NY.

Qureshi, S.N., M. Altaf-Khan, J.N. Jenkins, C.L. Brubaker, U.K. Reddy & S. Saha, 2001. A genetic linkage map of SSR markers based on *Gossypium nelsonii* and *G. australae* F_2 population. Plant and Animal Genome Meeting IX, Jan. 13–17. San Diego, California. Abst., pp. 219.

Ramsey, L., M. Macaulay, L. Cardle, M. Morgante, S. Ivanissevich, E. Maestri, W. Powell & R. Waugh, 1999. Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J 17: 415–425.

Reddy, O.U., A.E. Pepper, I. Abdurakmonov, S. Saha, J. Jenkins, T. Brook & M. El-Zik, 2001. New dinucleotide and trinucleotide microsatellite marker resources for cotton genome research. J Cotton Sci 5: 103–113.

Saghai-Maroof, M.A., R.M. Biyashev, R.M. Yang, G.P. Zhang & R.W. Allard, 1994. Extraordinarily polymorphic microsatellite DNA in barley: Species diversity, chromosomal locations & population dynamics. Proc Natl Acad Sci 91: 5446–5470.

Saha, S., M. Karaca, J.N. Jenkins & D. Lang, 2001. Improved methods for screening cotton cDNAs using flurochrome-labeled AFLP- and SSR-specific primers. In: J.N. Jenkins & S. Saha

(Eds.), Genetic Improvement of Cotton: Emerging Technologies, pp. 239–255. Science Publishers, Inc., Enfield, NH.

Saha, S., H. Tan, M. Karaca & J.N. Jenkins, 1999. Gel based DNA marker technologies in cotton. ICAR 17: 8–11.

Song, P., E. Yamamoto & R.D. Allan, 1995. Improved procedure of differential display of transcripts from cotton tissue. Plant Mol Biol Rptr 13: 175–181.

Wren, D.J., E. Forgacs, J.W. Fondon III, A. Pertsemlidis, Y.S. Cheng, T. Gallardo, R.S. Williams, R.V. Shohet, J.D. Minna & H.R. Garner, 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. Am J Hum Genet 67: 345–356.

Yu, K., S.J. Park & V. Poysa, 1999. Abundance and variation of microsatellite DNA sequences in beans (*Phaseolus* and *Vigna*). Genome 42: 27–34.